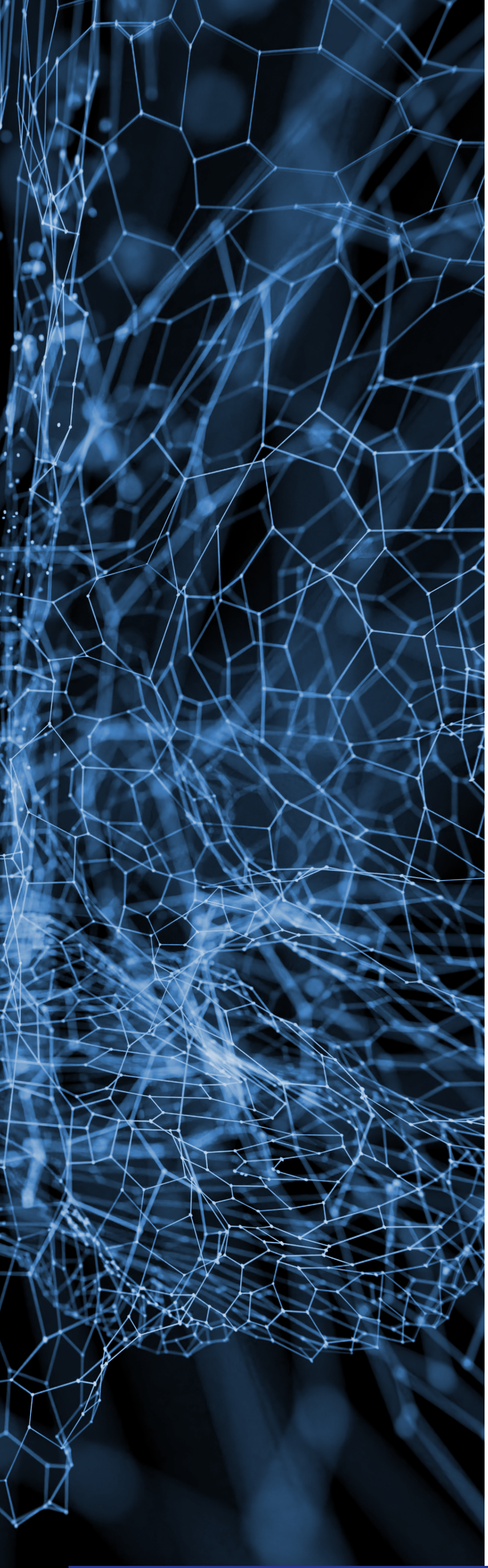


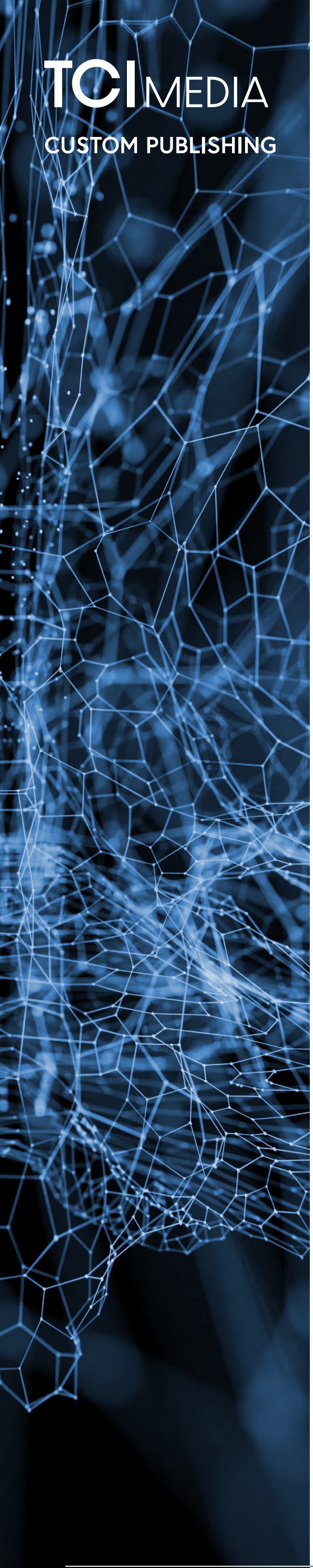
MODERN COST CONTROL

Chargeback for GPU and Hybrid Cloud Environments

Produced by TCI Media Custom Publishing in conjunction with:

HPCwire

Parallel Works



The shift to hybrid and multi-cloud infrastructure has transformed how organizations approach high-performance and AI workloads. Yet as this transformation accelerates, a growing challenge emerges: how to manage resource usage, justify infrastructure spend, and foster accountability across increasingly shared and more complex computing environments. For universities, government labs, defense agencies, and enterprises alike, critical compute resources are often poorly understood from a cost and consumption standpoint.

At the same time, organizations are making large investments in computing infrastructure as they move from AI SaaS and external service providers to internal and private AI systems. These systems are generally large investments and thus shared across an organization, and need to be charged back to those using the systems accurately and fairly.

Unfortunately, in most research and engineering environments today, HPC systems offer little to no built-in visibility into how resources are used or who is consuming them. On-prem environments often can track usage but lack unified tools; cloud billing is detailed but separate and often delayed. Additionally, cloud resources are elastic and decentralized, and the proliferation of environments, including Slurm, Kubernetes, VMware, and cloud platforms, means that a one-size-fits-all budgeting model simply doesn't work.

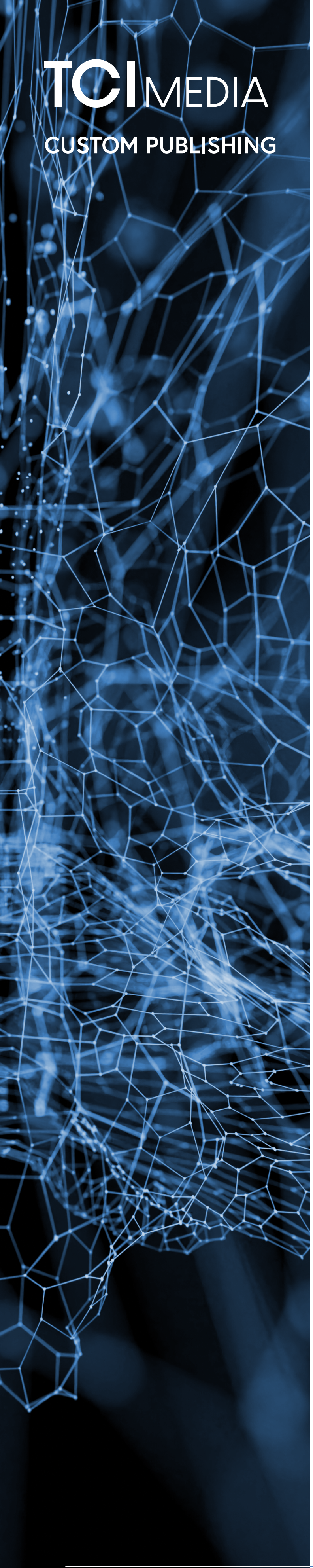
This is where ACTIVATE, the compute control plane from Parallel Works, comes into play. Designed to bring order to hybrid chaos, ACTIVATE enables seamless chargeback functionality across compute environments. By tying real-world costs to consumption, ACTIVATE empowers organizations to shift from allocation to accountability and turn shared compute into a governed, efficient, and transparent service. Additionally, ACTIVATE allows end users to self-serve resources within budget constraints. ACTIVATE helps support a model based on internal leasing of compute infrastructure to improve utilization.

Today in cloud and hybrid computing environments, infrastructure is once again shared, dynamic, and costly. As a result, chargeback is being brought back to the forefront of strategic IT and compute resource management.

The Accountability Gap in Modern Research Computing

A brief examination of the current state of research computing helps frame the need for a new approach to chargeback.

Modern research computing operates under an outdated cost-recovery model. Budget allocations still follow traditional top-down paradigms: annual or quarterly funds



are assigned to projects, groups, labs, or departments. Yet the nature of computing consumption has changed. It is now bursty, bottom-up, and shaped by dynamic project demands. Traditional HPC orchestration and scheduling portals offer job access but no built-in chargeback capabilities.

Adding to the challenge is the fact that most research and engineering organizations operate shared infrastructure with no financial accountability. Many are transitioning from managed AI services to private on-prem AI capabilities. Without precise metering, compute clusters suffer from a mismatch between funding levels and actual usage patterns. They are often overprovisioned and underutilized. Monthly batch logs and incomplete dashboards obscure idle time, which can account for a significant portion of GPU capacity. Researchers and developers queue for compute time that appears to be "free" but is practically unavailable due to hoarding or inefficient usage.

Cloud offerings (and costs) explode

Meanwhile, cloud costs spin up rapidly. While each hyperscale cloud service provider has methods of tracking and labeling usage, those offerings are specific to each provider. Access to that information is subject to the cloud provider's delayed billing model. All of that makes it difficult or impossible to enforce fixed budgets in the cloud.

To that point, usage is elastic, decentralized, and billed externally. Making matters worse, offerings to meet the ever-growing HPC and AI compute needs continue to grow. Hyperscalers continue to introduce new offerings, particularly instances based on the latest generation of GPUs.

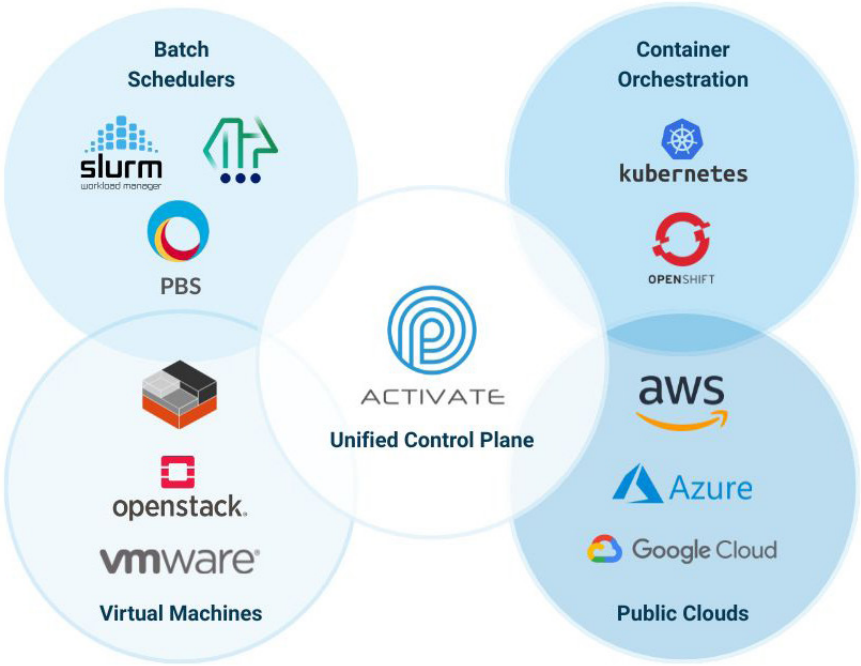
Researchers, developers, and other computing practitioners also have more choices with the emergence of new providers, dubbed neocloud providers, which offer specialized platforms featuring GPU-as-a-Service and AI-optimized infrastructure. Unlike traditional hyperscalers, neoclouds such as CoreWeave and Lambda offer on-demand access to high-performance GPUs with flexible pricing models tailored for AI training, inference, and research workloads.

Unlike traditional cloud hyperscalers, neoclouds such as CoreWeave and Lambda offer on-demand access to high-performance GPUs with flexible pricing models tailored for AI training, inference, and research workloads.

While this flexibility accelerates innovation, it introduces a new layer of financial and operational complexity in managing organizational computing costs, making it harder to enforce consistent metering, budgeting, and chargeback policies across environments. And as noted above, while hyperscalers do offer usage tracking and labeling tools, each cloud handles this differently and within a delayed billing cycle, making it difficult to enforce fixed budgets in real time.

The benefits and bane of heterogeneous compute environments

These challenges only grow more severe in heterogeneous environments. Organizations commonly run Slurm for batch jobs, Kubernetes for container orchestration, VMware clusters for virtual machines, and multiple public clouds for elastic workloads.



Each of these environments reports usage differently, if at all, and lacks a unified cost model. In the absence of consistent chargeback, teams resort to ad hoc scripts, piecemeal solutions, and delayed accounting.

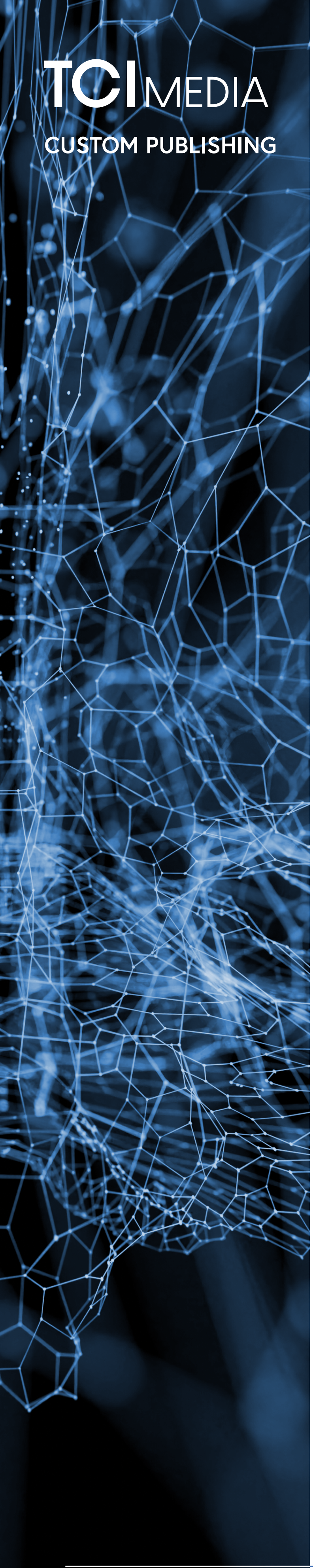
The result is a 'Wild West' of resource sharing, marked by overconsumption by some, underutilization by others, and a lack of trust in the system's fairness or efficiency.

Why Chargeback, and Why Now?

The reemergence of shared cluster chargeback models is being influenced strongly because GPU clusters are becoming more common in non-HPC organizations and, due to their high costs, operate in similar shared enterprise infrastructure modality as mainframes.

The return of chargeback reflects lessons from past computing models. Chargebacks played a crucial role in the era of mainframes, going back to the 1960s when large organizations centralized computing resources in IT departments. Mainframes were extremely expensive to purchase, operate, and maintain. Chargeback ensured fair cost allocation, with heavier compute users contributing a greater share.

Because compute time and storage were both limited and valuable, chargeback discouraged the submission of inefficient or unnecessary jobs. Without chargeback, users had no incentive to limit their use of mainframe resources, leading to waste and contention. With chargeback, users became more deliberate in submitting workloads, optimizing programs to use fewer CPU cycles or less memory, because they were effectively "paying" for it.



Additionally, chargeback systems gave organizations visibility into usage patterns and trends. This made spending more predictable, enabled better forecasting, and allowed departments to plan expenditures in line with their business priorities.

Chargeback reinforced the idea that computing was a shared but finite resource, not a free utility. This mindset shaped how enterprises approached compute consumption well into the era of client-server computing and early cloud models.

The principles that once justified chargeback in the mainframe era now apply to GPUs and organizational computing environments.

These issues are being revisited today in cloud and hybrid computing environments, where infrastructure is once again shared, dynamic, and costly. As a result, chargeback is being brought back to the forefront of strategic IT and compute resource management.

A Deeper Look at Modern Chargeback

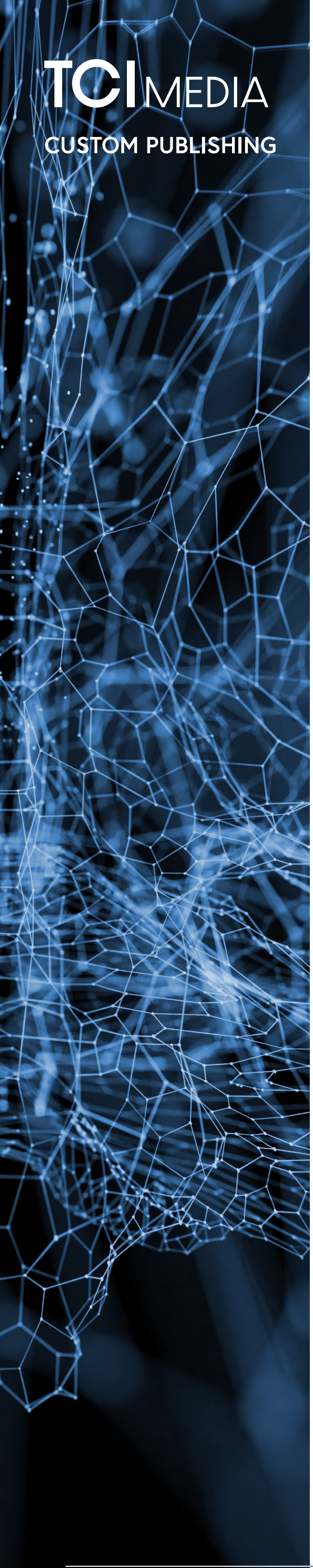
Chargeback is the practice of attributing compute usage to the users, projects, or departments that consume those resources, either through actual billing or symbolic 'showback' reporting. Chargeback is more than just a financial mechanism. At its best, chargeback serves as an incentive system, encouraging responsible usage and discouraging wasteful habits.

Much like the days of mainframe billing, modern chargeback drives efficiency via cost attribution in compute environments. It pushes researchers to optimize their workflows, right-size their jobs, and plan their budgets more effectively.

Specifically, in the context of research computing, the case for chargeback has never been more urgent. GPU clusters now routinely cost between $100,000 and $1 million, yet many organizations report utilization rates of less than 40%, according to industry sources and studies.

Low utilization rates in computing environments signal that expensive infrastructure is sitting idle for much of the time. This inefficiency results in higher costs per unit of work, meaning organizations pay for capacity they aren't actively using. Over time, this leads to a poor return on investment, unnecessary operational overhead, and delays in project execution due to resource bottlenecks in some areas while others remain underused. Furthermore, fragmented usage across environments (e.g., batch jobs on Slurm, containers on Kubernetes, virtual machines on VMware, and cloud-based workloads) can compound the problem if workloads aren't intelligently orchestrated across these platforms.

Raising utilization rates is critical because it unlocks the full value of an organization's computing assets. Higher utilization means more work gets done with existing infrastructure, allowing teams to scale research, development, and analytics without



needing additional capital investment. It also strengthens the case to CFOs and other stakeholders that IT resources are being used efficiently and strategically.. Most importantly, better utilization fosters innovation by removing artificial resource constraints, accelerating experimentation, and supporting a broader range of workloads across the organization.

Battling Low Utilization with Chargebacks

Without cost visibility, those consuming computing resources often default to requesting additional nodes, even when this may not be the most efficient or cost-effective approach. That lowers utilization rates throughout an organization.

In the context of high-performance computing (HPC) and AI environments, chargeback is the process of allocating IT costs to the users, projects, or departments that consume computing resources, creating a direct line of accountability between resource usage and financial responsibility.

It is increasingly necessary because infrastructure is expensive, demand is unpredictable, and resources are often shared. Chargeback serves as a vital mechanism for restoring transparency and operational control over resource consumption. It shifts the perception of compute from an abstract utility to a measurable, trackable asset. By attributing costs to core-hours, GPU-seconds, or terabyte-months of storage, organizations empower users to make more informed decisions about when, where, and how they run their workloads.

A key aspect of modern chargeback is enabling an end user self-service infrastructure within budget constraints—right-sizing workloads and improving overall system utilization. This drives significant savings by reducing idle resource waste and helps organizations reach utilization and productivity targets faster after large infrastructure investments.

Such visibility has wide-ranging benefits. Chargeback:

- Fosters a behavioral change in the way researchers and developers use computing resources, leading to more responsible use, avoiding unnecessary jobs, or optimizing their code to reduce runtime.
- Plays a critical role in budgeting and planning, helping departments forecast their needs based on historical consumption patterns.
- Provides leadership with data to justify future infrastructure investments or reallocate resources where they'll have the most impact.

Perhaps most importantly in academic and government settings, chargeback supports grant compliance by ensuring that compute costs are traceable to specific projects or funding sources.

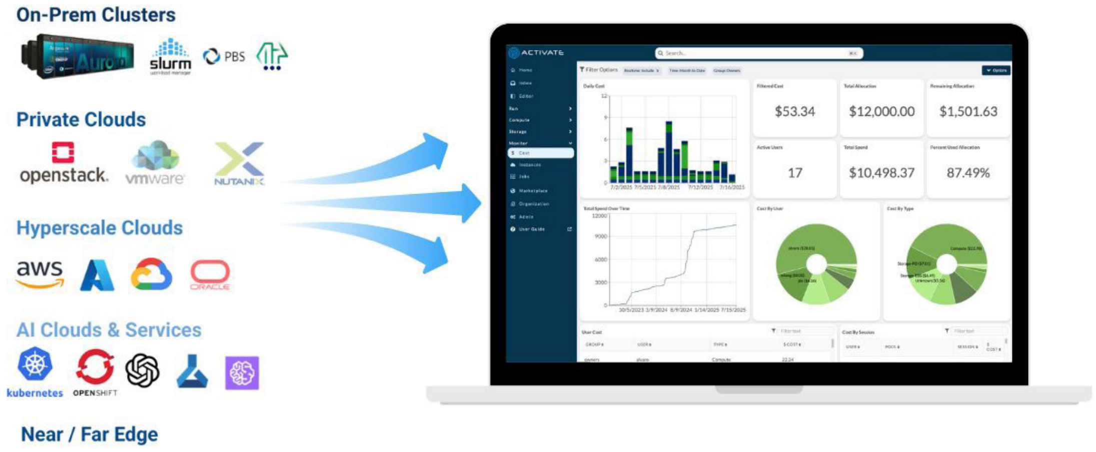
To that point, as organizations move away from flat, subsidized models, chargeback provides a bridge to more sustainable, consumption-based approaches that align usage with value. As a result, chargeback is increasingly expected in modern, cloud-based compute environments. That is especially the case in higher ed, government, and other HPC-dominated fields.

Early adopters of chargeback mechanisms often report improvements. The Uptime Institute reports that as users become aware of the direct costs of their activities, they adopt more efficient behaviors, leading to improved utilization, reduced waste, and deferment of infrastructure upgrades. The Institute further notes that users are motivated to optimize usage and eliminate inefficiency once their compute consumption is visible and made financially meaningful.

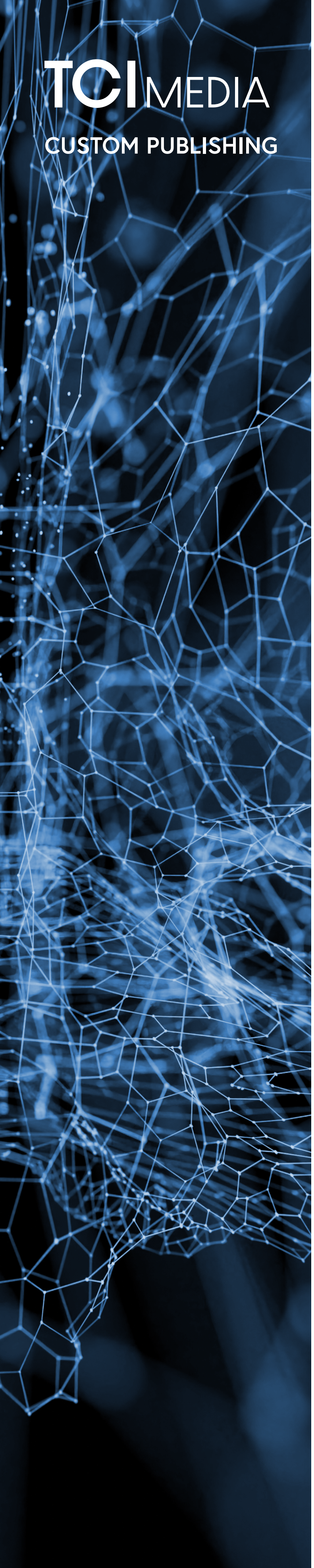
The ACTIVATE Approach

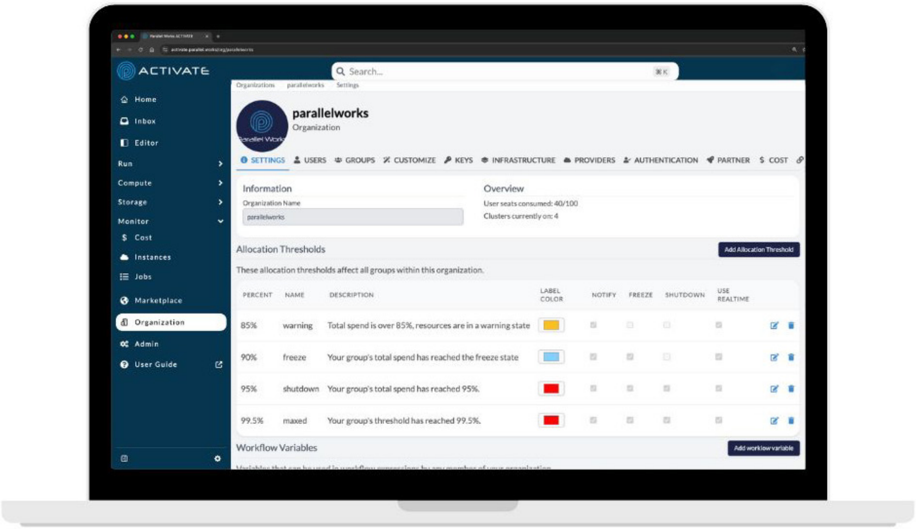
Parallel Works built ACTIVATE to address the challenges in hybrid and multicloud research environments. At its core, ACTIVATE offers a unified control plane that treats compute like any other critical resource: metered, monitored, and governed according to policy.

What makes ACTIVATE effective is its broad coverage, tight integration across environments, and user-transparent design. Rather than requiring researchers to change how they work, ACTIVATE integrates behind the scenes, plugging into batch schedulers like Slurm, orchestrators like Kubernetes, cloud billing APIs, and legacy VM environments.



By aggregating all this telemetry into a centralized cost-tracking framework, ACTIVATE allows organizations to apply consistent policies and price models across diverse systems. Whether a job runs on a bare-metal node, in a Kubernetes pod, or on a burstable AWS GPU instance, ACTIVATE tracks the usage and calculates the corresponding cost.





ACTIVATE

Search... ⌘K

Organizations parallelworks Settings

Home
Inbox
Editor
Run
Compute
Storage
Monitor
$ Cost
Instances
Jobs
Marketplace
Organization
Admin
User Guide

parallelworks
Organization

SETTINGS | USERS | GROUPS | CUSTOMIZE | KEYS | INFRASTRUCTURE | PROVIDERS | AUTHENTICATION | PARTNER | $ COST

Information

Organization Name: parallelworks

Overview

User seats consumed: 40/100
Clusters currently on: 4

Allocation Thresholds

Add Allocation Threshold

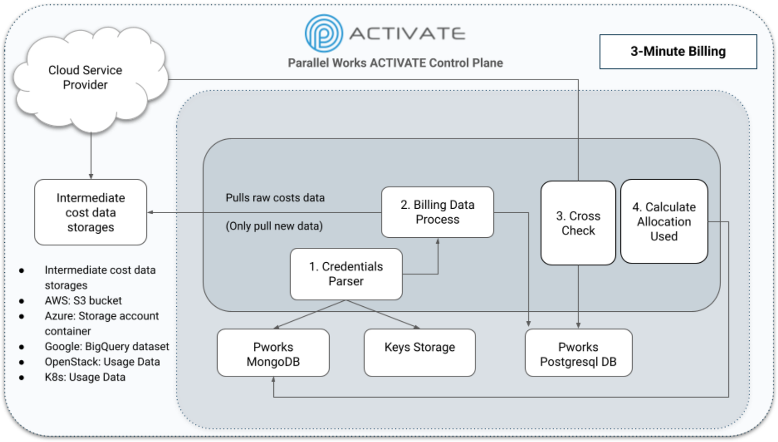
These allocation thresholds affect all groups within this organization.

PERCENT	NAME	DESCRIPTION	LABEL COLOR	NOTIFY	FREEZE	SHUTDOWN	USE REALTIME
85%	warning	Total spend is over 85%, resources are in a warning state	🟨	☑	□	□	☑
90%	freeze	Your group's total spend has reached the freeze state	🟦	☑	☑	□	☑
95%	shutdown	Your group's total spend has reached 95%.	🟥	☑	☑	☑	☑
99.5%	maxed	Your group's threshold has reached 99.5%.	🟥	☑	☑	☑	☑

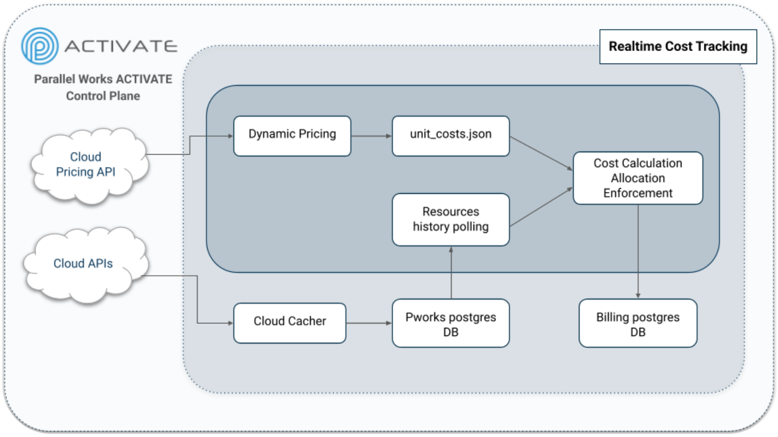
Workflow Variables

Add workflow variable

Administrators can define access policies, quotas, and budgets tied to specific users, groups, departments, or cost centers. Job attributes, such as runtime, job type, or resource class, can influence how costs are allocated or flagged. ACTIVATE also provides dashboards and reports that surface insights in real time, from budget burn rates to idle resource detection.



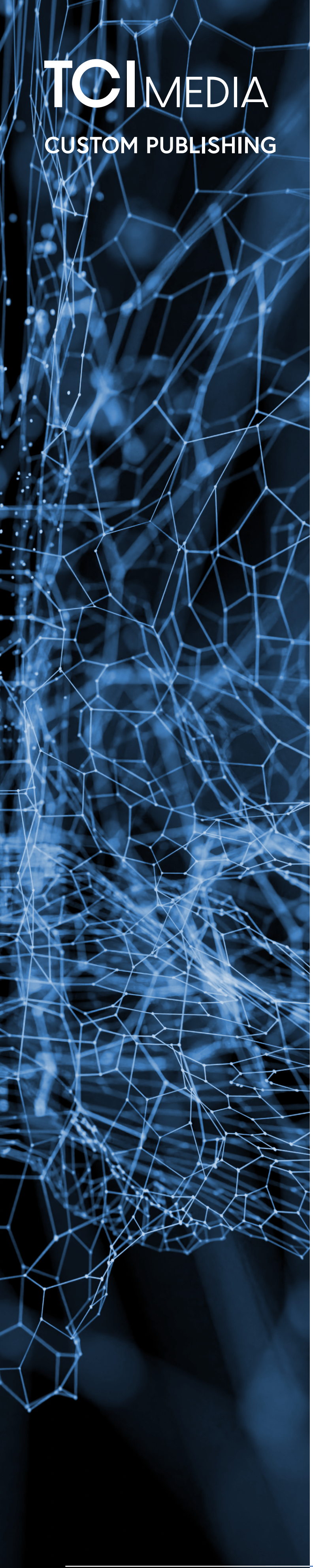
ACTIVATE supports both "showback" and full chargeback modes. In showback, users view their usage and associated costs, but no actual billing occurs. This mode helps organizations build trust, test their models, and identify inefficiencies before transitioning to full chargeback.



Use Cases in the Field

Across sectors, the need for chargeback is growing, and ACTIVATE's flexible architecture supports a range of sector-specific requirements.

In university research computing environments, ACTIVATE helps allocate compute hours to faculty grants and academic departments. This ensures fair access to shared clusters and supports grant compliance reporting. (Learn more about how Albert Einstein College of Medicine is using ACTIVATE.)



In the pharmaceutical industry, GPU-intensive AI pipelines can be tracked by project or therapeutic area, helping teams quantify the cost of each training run and prioritize accordingly. (See how pharmaceutical research benefits from ACTIVATE.)

In the energy sector, where large-scale simulations involve multiple teams and models, ACTIVATE enables organizations to bill usage based on actual consumption, rather than relying on guesswork. (Explore the use of ACTIVATE in the energy field.)

In government or defense environments where compliance is paramount, ACTIVATE supports budget traceability at Controlled Unclassified Information (CUI) levels (including IL5), enabling secure and accountable operations at scale. (See how ACTIVATE supports mission-critical HPC and AI government workloads.)

Governance Without Friction

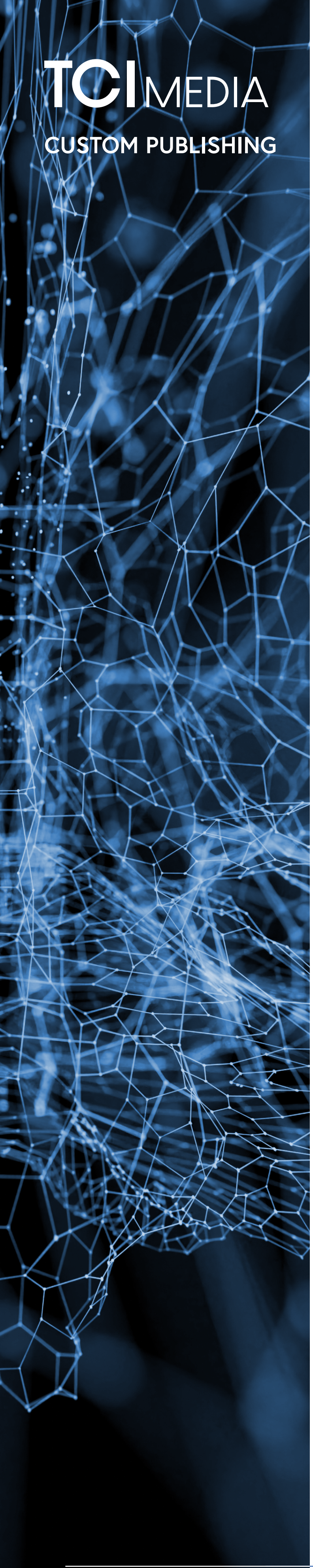
One of ACTIVATE's most compelling qualities is its ability to introduce governance without disrupting user workflows. Researchers are not expected to manually tag jobs or learn new tools. Instead, ACTIVATE automates attribution through integrations with identity management systems and job schedulers.

Such seamlessness is critical to driving adoption. Researchers, developers, and other computing practitioners gain visibility into their usage without feeling burdened. Administrators gain actionable data without having to stitch together logs. And leadership gains the cost accountability they need to make informed strategic decisions.

More broadly, ACTIVATE facilitates a cultural shift. When users understand the cost of their compute choices, their behavior changes. Instead of defaulting to the largest available resource allocation or rerunning failed experiments out of habit, teams begin to ask: "What's the smartest way to run this?"

This shift boosts efficiency and increases the impact of every dollar spent on infrastructure.

> ***"Chargeback fundamentally changed how our teams think about compute. Utilization jumped, idle GPUs became a thing of the past, and researchers began asking smarter questions about how to maximize their performance. It wasn't a penalty — it was a wake-up call." — ACTIVATE Customer***



Looking Ahead: The Future of Chargeback

As AI workloads become increasingly demanding and compute environments become even more hybridized, the need for intelligent, automated chargeback will intensify. The future will bring even more granular metering, such as per-second GPU billing, and smarter forecasting through machine learning and Token-based allocations across an organization for public or private LLMs and agentic services.

ACTIVATE is already evolving in this direction. Its integration with NVIDIA's NVML libraries and Slurm plug-ins enables second-level GPU monitoring. Its policy engine can also incorporate sustainability metrics—such as carbon intensity or local energy prices—enabling chargeback strategies that factor in environmental impact.

Organizations that embrace these capabilities are better positioned to scale responsibly. They can justify infrastructure investments with hard data, identify underperforming systems, and balance equity with efficiency through intelligent subsidy and burst models.

As one ACTIVATE customer put it: "Chargeback fundamentally changed how our teams think about compute. Utilization jumped, idle GPUs became a thing of the past, and researchers began asking smarter questions about how to maximize their performance. It wasn't a penalty—it was a wake-up call."

Chargeback is a Strategic Advantage

In the age of AI, computing is a strategic asset. As such, it demands the same level of financial scrutiny, operational efficiency, and governance as any other critical enterprise resource.

ACTIVATE makes this possible. It delivers the tooling, visibility, and control that research computing environments need to move beyond outdated allocation models. Whether managing a campus-wide HPC cluster, spinning up cloud GPU fleets, or balancing a mosaic of on-prem, containerized, and virtualized environments, ACTIVATE helps organizations make sense of it all and optimize accordingly.

The benefits are clear: higher utilization, lower waste, faster time-to-insight, and a culture of responsible innovation.